Entanglement: Embarrassingly-scalable graphs

Keith Flanagan, Allyson L. Lister, Matthew Pocock, Jennifer Hallinan, Caroline Relton and Anil Wipat
1 School of Computing Science, Newcastle University, Newcastle upon Tyne, NE1 7RU, UK,
2 Institute of Genetic Medicine, Newcastle University, Newcastle upon Tyne, NE1 7RU, UK,
3 School of Social and Community Medicine, University of Bristol, Bristol, BS8 1PS, UK,
4 TAMH Ltd, UK

Introduction

The ARIES project (http://www.ariesepigenomics.org.uk/) analyses, integrates and presents epigenetic information for a range of human tissues using both Illumina Infinium 450k and BS-seq technologies. Epigenetic data from the Avon Longitudinal Study of Parents and their Children cohort (ALSPAC, http://www.bristol.ac.uk/alspac/) is combined with several other publicly-available resources to form a integrated dataset.

The properties of graphs make them ideal for use in data integration applications. Data from multiple sources can be stored and interconnected within a graph, enabling data to be queried across different datasets. For ARIES, we require a scalable graph storage system capable of manipulating hundreds of millions to billions of graph entities. Entanglement is a software platform that provides a generic, scalable, graph framework suitable for data integration applications that require embarrassing scalability.

Scalability:
Billions of indexed graph entities partitioned over a cluster of machines.

Provenance and version control:
Every change to every node/edge is tracked over time.

Multiple independent graphs:
Datasets from different sources, or from different analyses can be placed in their own independent graphs.

Flexible, modular graph imports:
Large, integrated datasets can be composed by importing from several smaller graphs.

Entanglement focuses on features that are necessary for large-scale, graph-based data integration tasks.

Entanglement

Architecture

Entanglement uses MongoDB distributed document database to transparently distribute data and querying over a cluster of machines. An API provides an access to a graph and its revision history. Every version of each graph entity is stored and can be tracked over time. Domain-specific node and edge types, such as 'Chromosome', 'Gene', and 'located-on' are then defined. Parsers for project-specific file formats and graph queries can then be constructed.

Composition of integrated graphs

Although the ultimate aim is a fully integrated dataset, by intentionally storing different data sources in different graphs a large amount of flexibility can be obtained. Multiple ad-hoc integrated views can be composed by importing references to the nodes and edges in various individual dataset graphs.

Scalability

Entanglement permits multiple graphs to be populated in parallel. In contrast with replicated relational data stores, a cluster of MongoDB servers has a much higher write throughput than a single machine. Entanglement makes use of this by striping graphs across a MongoDB cluster, thereby allowing domain specific distributed applications to populate several graphs simultaneously.

There is no inherent limit on the number of graphs supported by Entanglement. Our system is suitable for use-cases where the results of each job in a distributed compute cluster could be stored in a separate graph for later integration. This architecture enables embarrassingly-parallel applications by removing almost all need for cluster-wide locks.

Data browsing

An interactive shell environment that allows users to query graphs and step through the nodes and edges, in much the same manner as a text-based Multi-User Dungeon (MUD).

Summary

We have developed Entanglement in order to provide embarrassingly-scalable graph storage. The framework supports a full provenance trail for every graph entity, allowing the user to know exactly which versions of data sources were used for a given analysis. Additional utilities have been written to allow Entanglement to integrate with Gephi and ONDEX.

Entanglement encourages the use of multiple independent graphs which can populated in parallel and then be recombined to form different integrated views over the available datasets.

Entanglement is released under the Apache 2.0 license. The system is under active development and we would welcome comments, suggestions and collaboration.

http://www.entanglementgraph.com